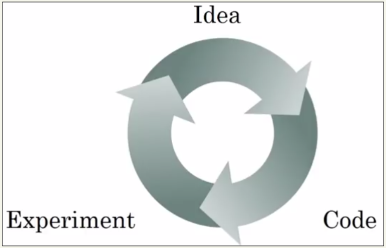
Aprendizaje no Supervisado

Introducción a la Inteligencia Artificial

Agenda

Aprendizaje Estadístico

- No se conoce la verdadera estadística.
- Se aprende por medio de datos.
- El buen desempeño no debe limitarse a los datos conocidos.



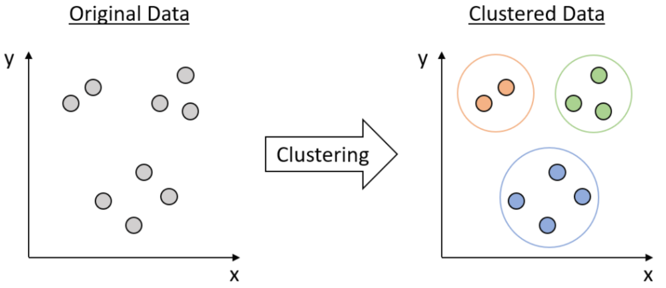
TIPOS DE APRENDIZAJES

- Aprendizaje supervisado: Cuento con pares de datos $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$.
- Aprendizaje no supervisado: Cuento solamente con datos $\{\mathbf{x}^{(i)}\}_{i=1}^{n}$.
- Aprendizaje semi-supervisado: Cuento con muchos datos no supervisados y unos pocos supervisados.

Clustering

Clustering

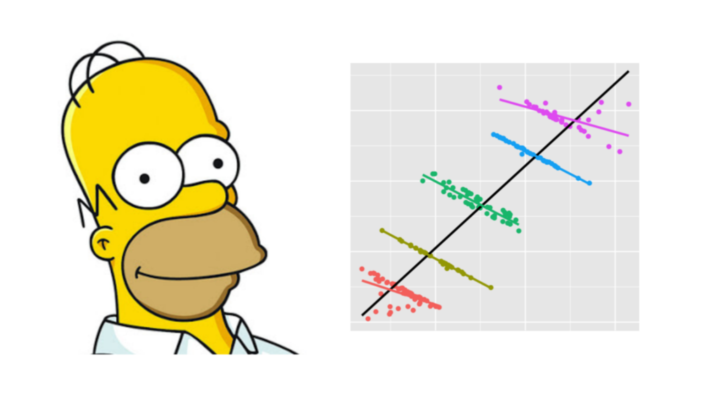
Estos algoritmos son la versión no supervisada de la clasificación. Su objetivo es agrupar muestras de manera de tener un mayor entendimiento sobre su naturaleza.



Motivación: Paradoja de Simpson

Paradoja de Simpson

La paradoja de Simpson se da cuando dos (o más) variables tienen una correlación hacia un sentido pero al agrupar los datos se ve que, en cada cluster, la correlación posee en realidad el sentido opuesto.



Paradoja de Simpson: Covid-19 Case Fatality Rates (CFR)

Edad	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	$\geq$ 80	Total
Italia	0% (0/43)	0% (0/85)	0% (0/296)	0% (0/470)	0.1% (1/891)	0.2% (3/1453)	2.5% (37/1471)	6.4% (114/1785)	13.2% (202/1532)	4.4% (357/8026)
China	0% (0/0)	0.2% (1/549)	0.2% (7/3619)	0.2% (18/7600)	0.4% (38/8571)	1.3% (130/10008)	3.6% (309/8583)	8% (312/3918)	14.8% (208/1408)	2.3% (1023/44672)

Julius von Kugelgen, Luigi Gresele and Bernhard Scholkopf "Simpson's paradox in Covid-19 case fatality rates: A mediation analysis of age-related causal effects" IEEE Transactions on Artificial Intelligence 2021.

Algoritmo K-Means

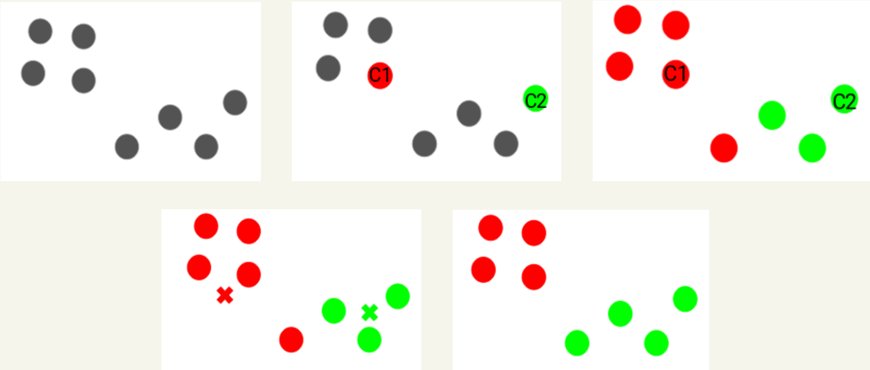
K-means

Algoritmo de clustering para agrupar los datos en K clusters (previamente definidos). Se basa en encontrar, de forma iterativa, los *centroides* de cada clase y asignar cada muestra al centroide más cercano.

Algorithm 1 K-means

```
procedure KMEANS(X, K)
Input: X ∈ ℝ^{n×d_x} matriz de datos y K número de clusters.
Output: μ ∈ ℝ^{K×d_x} centroides e y ∈ {1, ··· , K}^n etiquetas.
    Inicializar μ con el valor de K columnas de X elegida al azar.
    repeat
        y[i] = arg min_k ‖X[i, :] − μ[k, :]‖          ▷ Con i = 1, ··· , n.
        μ[k, :] = 𝔼 [X[y == k, :]]                   ▷ Con k = 1, ··· , K
    until convergencia
    Return: μ e y
end procedure
```

Algoritmo K-Means



Outline

Máxima Verosimilitud

Algoritmos de Máxima Verosimilitud

La minimización de la *cross-entropy* equivale a encontrar algoritmos de máxima verosimilitud. El problema es que estos son muchas veces analíticamente intratables y computacionalmente muy pesados de tratar (ej. mezcla de gaussianas).

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p(X_i|\theta)$$

Variables no observable

Sea Z una variable no observable del problema con densidad condicional $p(z|x,\theta)$, y sea $\mathcal{P}$ la familia de todas las posibles densidades condicionales de $Z|X=x$. Luego, el estimador de MV puede reescribirse como:

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta \in \Theta} \max_{q \in \mathcal{P}} \sum_{i=1}^{n} \left[\log p(X_i|\theta) - \mathrm{KL}\left(q(\cdot|X_i) \| p(\cdot|X_i,\theta)\right)\right] \\ &= \arg\max_{\theta \in \Theta} \max_{q \in \mathcal{P}} \mathrm{ELBO}(\theta, q)\end{aligned}$$

Algoritmo EM

Algoritmo Expectation - Maximization

El algoritmo EM consiste en inicializar en algún valor θ_0 e iterar entre:

- $q^{(t)} = \arg\max_{q \in \mathcal{P}} \text{ELBO}(\theta^{(t-1)}, q) \quad \text{(Expectation)}$
- $\theta^{(t)} = \arg\max_{\theta \in \Theta} \text{ELBO}(\theta, q^{(t)}) \quad \text{(Maximization)}$

Algoritmo EM

Algoritmo Expectation - Maximization

El algoritmo EM consiste en inicializar en algún valor θ_0 e iterar entre:

- $q^{(t)} = \arg\max_{q \in \mathcal{P}} \ \mathrm{ELBO}(\theta^{(t-1)}, q) \quad \text{(Expectation)}$
- $\theta^{(t)} = \arg\max_{\theta \in \Theta} \ \mathrm{ELBO}(\theta, q^{(t)}) \quad \text{(Maximization)}$

Expectación

El paso *Expectation* puede simplificarse a la relación $q^{(t)}(z|x) = p(z|x, \theta^{(t-1)})$. Es decir:

$$\theta^{(t)} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \left[\log p(X_i|\theta) - \mathrm{KL}\left(p(\cdot|X_i, \theta^{(t-1)}) \| p(\cdot|X_i, \theta) \right) \right]$$

Algoritmo EM

Maximización

El paso *Maximization* puede simplificarse reesribiendo cada sumando como

$$\log p(x|\theta) - \mathrm{KL}\left(q(\cdot|x) \| p(\cdot|x,\theta)\right) = H(q(\cdot|x)) + \mathbb{E}_q\left[\log p(x, Z|\theta) | X = x\right]$$

donde la entropía no depende de θ. Es decir,

$$\theta^{(t)} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \mathbb{E}_{q^{(t)}}\left[\log p(X_i, Z|\theta) | X_i\right]$$

Algoritmo EM

Maximización

El paso *Maximization* puede simplificarse reesribiendo cada sumando como

$$\log p(x|\theta) - \mathrm{KL}\left(q(\cdot|x)\|p(\cdot|x,\theta)\right) = H(q(\cdot|x)) + \mathbb{E}_q\left[\log p(x,Z|\theta)|X=x\right]$$

donde la entropía no depende de θ. Es decir,

$$\theta^{(t)} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \mathbb{E}_{q^{(t)}}\left[\log p(X_i,Z|\theta)|X_i\right]$$

Teorema: Monotonía

En el algoritmo EM ocurre que

$$\sum_{i=1}^{n} \log p\left(X_i|\theta^{(t)}\right) \geq \sum_{i=1}^{n} \log p\left(X_i|\theta^{(t-1)}\right)$$

Hint: Expectación + KL ≥ 0.

Algoritmo EM para mezcla de gaussianas

Definición del problema

Si $Z \sim \text{Cat}(\{c_1, \cdots, c_K\})$ y $X|Z = k \sim \mathcal{N}(\mu_k, \Sigma_k)$, está claro que X es una mezcla de gaussianas. Sea $\theta = \{c_k, \mu_k, \Sigma_k\}_{k=1}^{K}$, se desea estimar estos parámetros (de forma no supervisada, es decir siendo Z no observable). El estimador de máxima verosimilitud es intratable y por eso recurrimos al algoritmo EM.

Algoritmo EM para mezcla de gaussianas

Definición del problema

Si $Z \sim \text{Cat}(\{c_1, \cdots, c_K\})$ y $X|Z = k \sim \mathcal{N}(\mu_k, \Sigma_k)$, está claro que X es una mezcla de gaussianas. Sea $\theta = \{c_k, \mu_k, \Sigma_k\}_{k=1}^{K}$, se desea estimar estos parámetros (de forma no supervisada, es decir siendo Z no observable). El estimador de máxima verosimilitud es intratable y por eso recurrimos al algoritmo EM.

Expectación

El paso de expectación es simplemente elegir:

$$q(k|x) = p(k|x, \theta) = \frac{c_k \cdot \left|\Sigma_k\right|^{-1/2} \cdot e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}}{\sum_{m=1}^{K} c_m \cdot \left|\Sigma_m\right|^{-1/2} \cdot e^{-\frac{1}{2}(x-\mu_m)^T \Sigma_m^{-1}(x-\mu_m)}}$$

Clustering

En los modelos de mezcla (cuando Z es una variable categórica), se considera al algoritmo EM un método de *soft clustering*.

Algoritmo EM para mezcla de gaussianas

Maximización

Dado un q, se desea maximizar:

$$\max_{\theta} \sum_{i=1}^{n} \mathbb{E}_q \left[\log p(X_i, Z|\theta)|X_i\right] \qquad \text{s.t.} \qquad \sum_{k=1}^{K} c_k = 1$$

Es decir que, utilizando multiplicadores de Lagrange, la función a derivar e igualar a cero es:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} q(k|x_i) \left[\log c_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\right] + \lambda \left(1 - \sum_{k=1}^{K} c_k\right)$$

Algoritmo EM para mezcla de gaussianas

Derivada respecto a c_k

Igualamos a cero la derivada respecto a c_k y usamos que $\sum_{k=1}^{K} c_k = 1$.

$$\frac{\partial \mathcal{L}(\theta)}{\partial c_k} = \left(\sum_{i=1}^{n} \frac{q(k|x_i)}{c_k} \right) - \lambda = 0$$

$$\Rightarrow \quad c_k = \frac{1}{\lambda} \sum_{i=1}^{n} q(k|x_i)$$

$$\Rightarrow \quad c_k = \frac{1}{n} \sum_{i=1}^{n} q(k|x_i)$$

Algoritmo EM para mezcla de gaussianas

Derivada respecto a μ_k

Igualamos a cero (vector) la derivada respecto a μ_k.

$$\frac{\partial \mathcal{L}(\theta)}{\partial \mu_k} = \sum_{i=1}^{n} q(k|x_i)\Sigma_k^{-1}(x_i - \mu_k) = 0$$

$$\Rightarrow \quad \mu_k = \frac{\sum_{i=1}^{n} q(k|x_i) \cdot x_i}{\sum_{i=1}^{n} q(k|x_i)}$$

Petersen and Pedersen - "Matrix Cookbook".

Algoritmo EM para mezcla de gaussianas

Derivada respecto a μ_k

Igualamos a cero (vector) la derivada respecto a μ_k.

$$\frac{\partial \mathcal{L}(\theta)}{\partial \mu_k} = \sum_{i=1}^{n} q(k|x_i)\Sigma_k^{-1}(x_i - \mu_k) = 0$$

$$\Rightarrow \quad \mu_k = \frac{\sum_{i=1}^{n} q(k|x_i) \cdot x_i}{\sum_{i=1}^{n} q(k|x_i)}$$

Derivada respecto a Σ_k

Igualamos a cero (matriz) la derivada respecto a Σ_k.

$$\frac{\partial \mathcal{L}(\theta)}{\partial \Sigma_k} = \sum_{i=1}^{n} q(k|x_i) \left[-\frac{1}{2}\Sigma_k^{-1} + \frac{1}{2}\Sigma_k^{-1}(x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1} \right] = 0$$

$$\Rightarrow \quad \Sigma_k = \frac{\sum_{i=1}^{n} q(k|x_i) \cdot (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} q(k|x_i)}$$

Petersen and Pedersen - "Matrix Cookbook".